Wrangle Report

Wrangling contains usually three parts

- Gathering – can be different sources like databases web page scraping or using APIs
- Accessing – Getting to know your data and the difficulties in terms of tidiness and quality issues. List them
- Cleaning – Following your finding from the accessment and define the necessary steps to solve them in an appropriate way

Gathering

We had three sources of data to gather.

- twitter-archive-enhanced
- predictionData
- tweet-json.txt

First goal was to store them all in pandas data frame format.

For the Twitter archive that was easy since it was given as .csv flat text file. Just using plain pandas and read_csv method.

The prediction data required to apply the request libraries for getting data from a given URL. After download I used again read_csv method to assign it into a data frame.

No the a bit more complicate part. The Tweet json.txt file a specific pattern so that each line was its own json object and that made it confusing in the first step. After some tests I found two possibilities to handle this and put the json objects for each tweet into a pandas data frame finally

1. see cell 7 – more lines of code and a for loop
2. see cell 8 -- more elegant from my point of view

Having this done we have all three data source successfully stored into data frames. Ready for Accessing them.

Assessing

When assessing, you're like a detective at work, inspecting your dataset for two things: data quality issues (*i.e. content issues*) and lack of tidiness (*i.e. structural issues*).

For each of the three sources I applied a similar pattern using the following methods

- .head
- .tail
- .sample
- .info
- .describe
- .value_counts

I documented each observation as a guideline and clustered the topics into these two categories:

- Quality: issues with content. Low quality data is also known as dirty data.
- Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:
    a. Each variable forms a column.
    b. Each observation forms a row.
    c. Each type of observational unit forms a table.

I used two types of assessment:

- Visual assessment: I investigated the printed data by scrolling through
- Programmatic assessment: using code to view specific portions and summaries of the like the

Cleaning

I followed the different cleaning steps in best practice order as suggested by the course guides

1. Find Completeness Issues
    a. Find the missing data, create it and then concatenate, join or input. Sometimes one can't find it.
2. Cleaning for tidiness is usually the second step. Tidy data is easy to manipulate and so this should be done first.
    a. First Structural issues
    b. Then quality issues
3. Quality issues
    a. Everything remaining :-)

For cleaning the data I used the programmatical way and applied various method like:

- Dropna
- Replace with NaN
- Str.extract
- Merge
- Drop
- Todaytime
- Str is lower
- Str is capitalized
- Etc…

It helped me to follow the steps first missing data – then tidiness- then rest of quality issues.